Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 4298–4302, May 1994
Evolution

Codon reiteration and the evolution of proteins

HOWARD GREEN AND NORMAN WANG

Department of Cell Biology, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115

Contributed by Howard Green, January 18, 1994

ABSTRACT **Sequence data banks have been searched for proteins possessing uninterrupted reiterations of any amino acid. Hydrophilic amino acids, and particularly glutamine, account for a large proportion of the longer reiterants. In the genes for these proteins, the most common reiterants are those that contain poly(CAG), even out-of-frame or, to a lesser degree, those that contain repeated doublets of CA, AG, or GC. The preferential generation of such reiterants requires that DNA strand-specific signals predispose to reiteration and thus to the extension of coding regions.**

Sequences consisting of multiple consecutive residues of a single amino acid (reiterants) are not rare in proteins. For example, reiterants consisting of glutamine residues, most often encoded by CAG, were first discovered in homeotic proteins (1–3) and later in other transcription factors. In the case of the gene for involucrin, CAG reiteration seems to have been the means by which the coding region was generated (4–6). Human genetic diseases due to triplet reiteration are the result of uncontrolled action of the reiteration mechanism used in evolution (7).

To see what rules govern the reiteration process, we have analyzed the available data on the amino acid reiterants that have been introduced into proteins and on the nature of the responsible codons. We describe here the preferential reiteration of trinucleotides resembling the glutamine codon CAG.

Incidence of Amino Acid Reiterants of Different Lengths

A search was made of three sequence data banks (GenBank, GenBank Update, and the Brookhaven Protein Data Bank) using BLAST (8) for sequences consisting of uninterrupted amino acid reiterants of different length. The frequency of reiterants containing five to nine identical amino acid residues is shown in Fig. 1*A*. In general, the frequency of reiterants of each amino acid correlates with its abundance in proteins: leucine and alanine, the most abundant, are the most frequently reiterated, whereas tryptophan and cysteine, the least abundant, are virtually not reiterated. This suggests that for some amino acids, the principal factor affecting the frequency of reiterants of five to nine residues is the abundance of the amino acid; but even for this length of reiterant, there are amino acids, such as valine and isoleucine, that are underreiterated in relation to their abundance in proteins.

Reiterants containing 10–14 residues (Fig. 1*B*) are quite different from those containing 5–9 residues. Glutamine, which is not a very abundant amino acid, is the most frequently reiterated. Leucine, the most abundant amino acid, is infrequently reiterated, and there are no reiterants of valine, isoleucine, tyrosine, methionine, cysteine, and tryptophan. There are also no lysine reiterants.

Among reiterants containing 15–19 residues (Fig. 1*C*), glutamine is still the most frequent; phenylalanine has dropped out, leaving few reiterants of hydrophobic amino acids. There is no reiterant of arginine.

Among reiterants containing 20 or more residues, glutamine is clearly dominant and accounts for half of all reiterants. The other nine amino acid residues found in reiterants are mainly hydrophilic.

Codons Responsible for Reiterants of 10 or More Amino Acid Residues

Of the 229 reiterants, there are only 48 that are encoded by uninterrupted reiterations of a single codon; most of the amino acid reiterants are the result of mixed synonymous codons. Probably some of these were originally reiterants of a single codon, but nucleotide substitutions have since occurred within the sequence. In examining the frequency of codons responsible for amino acid reiterants, we have considered only the uninterrupted codon reiterants (Fig. 2). Of these, only 1 is in a prokaryote (*Mycoplasma*).

The reiterants may be divided into two classes according to whether the reiterated codon does or does not bear some resemblance to the codon CAG (Table 1). Of the total of 48 reiterants, CAG accounts for 15. Two codons, AGC (Ser) and GCA (Ala), are permutations of CAG and, when reiterated, give rise to out-of-frame reiterations of CAG; these two codons account for 6 reiterants.

The six serine reiterants are encoded by only two of its six synonymous codons. Of these two, only AGC, when reiterated, gives rise to out-of-frame poly(CAG), and this codon is responsible for five of the six examples of reiterated serine. This is so even though AGC is not particularly prominent among the serine codons used for a range of microbial, *Drosophila*, and human proteins (11). The only other serine codon found to be reiterated, TCA, contains the dinucleotide CA. Of the four nonreiterated serine codons, three lack any resemblance to CAG. Similarly, GCA is the only alanine codon (of four) whose reiteration gives rise to out-of-frame poly(CAG) and is the only codon responsible for an alanine reiterant. Accumulation of clusters of CAG and of permutations of that trinucleotide have been noted earlier in the gene for hunchback (*hb*) in one of two lineages of *Drosophila* (12) and in vertebrate and *Drosophila* genes for the TATA-binding protein (TBP) (13).

Of the remaining reiterated codons listed in Table 1, 20 possess the doublets CA, AG, or GC. Of the 10-codon reiterants that do not possess at least one of these dinucleotides, there are 6 examples of poly(AAT) (Asn), found only in microorganisms, and 1 example of poly(GAT) (Asp). If the rules of reiteration are somewhat different in microorganisms and this group is eliminated, there is left in animals and plants only a single reiterant of >10 codons lacking all of the three dinucleotides [poly(GAT)]. None of the reiterated codons are triplets of a single nucleotide.

A summary of the relation between the frequency of reiterants and the resemblance of their codons to CAG is given in Table 2. The three codons giving rise to reiterated

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: TBP, TATA-binding protein.

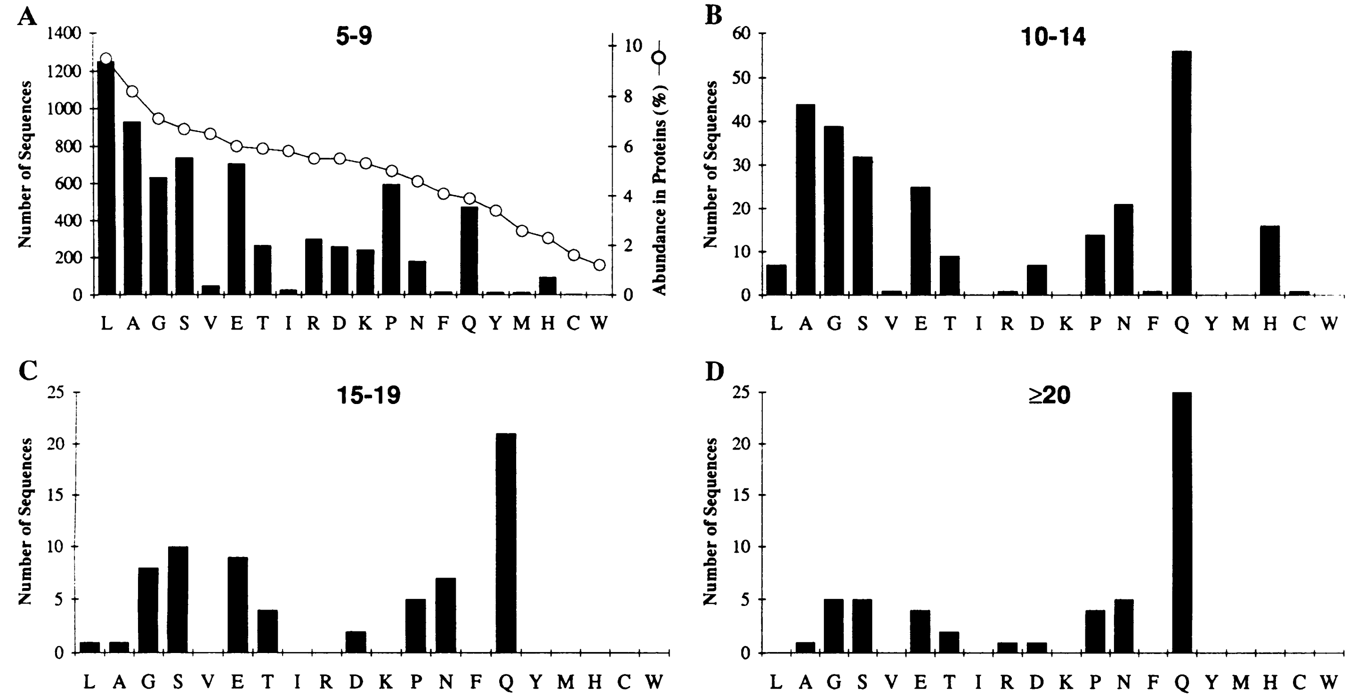

FIG. 1. Amino acid reiterants in proteins. The protein data banks were searched for uninterrupted reiterants of each amino acid. Each panel gives a different range of reiterant lengths. (*A*) Frequency of reiterants of five to nine identical amino acid residues, uncorrected for duplicate entries. Included is the abundance of each amino acid in a large number of proteins of molecular mass 20–50 kDa taken from ref. 9. For reiterants of greater length (*B*–*D*), duplicates have been removed. Note scale changes on ordinates.

CAG codons account for 21 of the reiterants or 44% of the total. These three codons have a frequency of reiteration per codon that is 30 times higher than codons lacking any resemblance to CAG. The reiteration frequency of codons containing two nucleotides of the three is intermediate, CA being most important and GC the least. If microbial forms were excluded from the comparisons, the ratio of reiterants containing CAG to reiterants lacking CA, AG, and GC would be much higher.

Multiple Codon Reiterants in a Single Gene

Near a reiterant in a coding region, there may be a second reiterant. For example, in the *Drosophila* notch gene (2) and in the LPMC61 antigen of *Eimeria tenella* (14), poly(CAG) stretches may be separated from each other by as little as one triplet that diverges from CAG by a single nucleotide. This suggests that the poly(CAG) stretches belong to a single longer sequence that was interrupted by a mutation. Other examples of synonymous and nonsynonymous codon interruptions are found in the polyglutamine sequences of the transcription factors TBP (15), SNF5 (16), and brahma (17).

Sometimes the nature of the interruption is far from random. For example, the amino acid sequence of the mastermind gene (*mam*) in *Drosophila virilis* (18) has a region (residues 106–161) containing 44 glutamine residues, of which 31 are encoded by CAG and 13 by CAA. All of the remaining 12 amino acid residues are histidines (encoded by CAC or CAT). These interrupt the glutamine codons at 10 locations and are likely to have resulted from single nucleotide substitutions in either CAG or CAA codons. This would seem to be both a high rate and a specific kind of divergence operating only on the third nucleotide of the glutamine codons.

In the *A26* and *A42* genes for α/β gliadin (19), there is a sequence of 12 uninterrupted CAA codons. In the *A42* gene, further CAA codons lie immediately upstream, but in *A26*, the adjacent upstream codons are CAG. In the case of *A26*, it seems that after a CAG codon was generated by mutation at the 5′ end of the reiterated CAA codons, this CAG, rather than the adjacent CAA codons, was reiterated.

Mixed Reiterations. In the glucose repression mediator SSN6 or CYC8 (20) of *Saccharomyces cerevisiae*, there is a stretch of 64 residues consisting almost entirely of alternating glutamine and alanine residues. In the encoding DNA, (GCA-CAA-GCA-CAA, etc.), 184 nucleotides (95%) of the total of 192 form CA, AG, or GC doublets. This sequence lies immediately 5′ of a sequence of 27 codons of almost homogeneous poly(CAA). It seems clear that near the 5′ end of this stretch, two events of substitution or of insertion/deletion led to the formation of the codon GCA. The pair of codons GCA-CAA was then reiterated, extending the alternation in the 5′ direction; the reiteration mechanism must have acted, not on a trinucleotide, but on the hexanucleotide or some multiple of it.

Low Frequency of T. It was pointed out earlier that the sense strand of the involucrin gene, whose evolution probably began by CAG reiteration, has a low T content (4). The reiterated codons listed in Table 1 have an average T content of 9%, whereas the unreiterated codons have a T content of 30%. None of the 10 codons containing more than one T is reiterated. The relatively low T content of reiterated codons may be necessary because a frameshift within a T-rich sequence could lead to termination.

Factors Affecting Length of Reiterant

The number of identical codons in a reiterant should depend on the activity of the reiterative mechanism, the rate of nucleotide substitution in the reiterated codons, the age of the reiterated sequence, and the selective pressures acting either against the protein containing the amino acid reiterant or against a codon reiterant that would cause difficulty in transcription or translation. It seems from this survey that reiterated hydrophobic or basic residues may be poorly tolerated in proteins, since there are practically no proteins containing 20 or more residues belonging to either category. In the fragile X mental retardation gene (21) and the myotonic dystrophy gene

Reiterant		Access.		
AA.	Codon	no.	Gene	Species
Ala	$(gca)_{11}$	M98269	antho-RFamide neuropeptide	A. elegantissima
Asn	$(aac)_{20}$	X16523	AAC-rich mRNA (pLK330)	D. discoideum
	$(aac)_{11}$	X54666	BRcore-TNT1-Q1-Z1 (Broad Complex)	D. melanogaster
	$(aat)_{11}$	X17487	asparagine-rich antigen	P. falciparium
	$(aat)_{11}$	M62622	var I, mitochondrial DNA	S. cerevisiae
	$(aat)_{16}$	S55235	cAR3=cAMP receptor subtype3	D. discoideum
	$(aat)_{11}$	X62147	nud1	S.cerevisiae
	$(aat)_{14}$	M87278	adenylyl cyclase germination protein	D. discoideum
	$(aat)_{11}$	X16561	RNA polymerase II large subunit	P. falciparum
Asp	$(gat)_{14}$	M60052	histidine-rich calcium binding protein	Homo sapiens
Gln	$(caa)_{11}$	L19349	hydroxymethylglutaryl CoA reductase	D. discoideum
	$(caa)_{15}$	M11074	α-/β-gliadin	T. aestivum
	$(caa)_{11}$	M76586	zinc finger protein	Candida albicans
	$(caa)_{20}$	M17826	SSN6 or CYC8	S.cerevisiae
	$(caa)_{15}$	M60807	merozoite surface antigen 1	P. falciparum
	$(caa)_{10}$	D10250	α-fetoprotein enhancer binding protein	Homo sapiens
	$(cag)_{12}$	X07422	interleukin 2	Mus musculus
	$(cag)_{10}$	M30933	antigen LPMC61	Eimeria tenella
	$(cag)_{21}$	L12392	Huntington's disease gene	Homo sapiens
	$(cag)_{10}$	L04487	castor zinc-finger protein	D. melanogaster
	$(cag)_{19}$	Y00489	ventral prostate glucocorticoid receptor	Rattus rattus
	$(cag)_{19}$	M55654	TATA-box binding protein (TBP)	Homo sapiens
	$(cag)_{15}$	X06832	prechromogranin	Rattus rattus
	$(cag)_{10}$	J04566	Vgr-1	Mus musculus
	$(cag)_{11}$	X68505	myocyte-specific enhancer factor 2	Homo sapiens
	$(cag)_{16}$	M23263	androgen receptor	Homo sapiens
	$(cag)_{11}$	M88300	brain-2 POU-domain protein	Mus musculus
	$(cag)_{14}$	M93690	RT1 retroposon	A. gambiae
	$(cag)_{14}$	L08424	achaete scute homologous protein	Homo sapiens
	$(cag)_{13}$	X72889	hbrm (SNF2/sw12 and brm homolog)	Homo sapiens
	$(cag)_{18}$	L28819	involucrin	Mus musculus
Glu	$(gaa)_{10}$	J03998	glutamic acid-rich protein	P. falciparum
	$(gaa)_{12}$	J03918	Sec7 gene (Golgi membrane control)	S.cerevisiae
	$(gag)_{12}$	J05080	histidine-rich calcium-binding protein	O. cuniculus
	$(gag)_{10}$	M18289	E1B small T-antigen	Adenovirus 41
Gly	$(gga)_{10}$	M18289	E1B large T-antigen	Adenovirus 41
	$(ggc)_{10}$	X04106	calcium-dependent protease	Homo sapiens
	$(ggc)_{20}$	M23263	androgen receptor	Homo sapiens
	$(ggc)_{10}$	D10250	α-fetoprotein enhancer binding protein	Homo sapiens
His	$(cat)_{11}$	S55234	cAMP receptor subtype 2	D. discoideum
Ser	$(agc)_{10}$	X15898	sporozite antigen	Eimeria tenella
	$(agc)_{11}$	L11275	SRP40	S,cerevisiae
	$(agc)_{12}$	J03149	c-fms proto-oncogene (M-CSF receptor)	Felis domesticus
	$(agc)_{10}$	M88749	vitellogenin	I. unicuspus
	$(agc)_{10}$	L13744	AF-9	Homo sapiens
	$(agt)_{11}$	M31431	attachment protein	Myco. genitalium
Thr	$(aca)_{17}$	M66619	aminocyclopropane carboxylate synth.	D. caryophyllus
	$(acc)_{12}$	S51097	intestinal alkaline phosphatase II	Rattus rattus

FIG. 2. Uninterrupted reiterants of ≥10 codons. For a summary of human genes containing trinucleotide repeats identified by less stringent criteria, see ref. 10.

(22–24), the length of their respective reiterants of CGG and of CTG, even in normal alleles, extends above the range of amino acid reiterations corresponding to these triplets in any frame (arginine, glycine, and alanine for the first and leucine, cysteine, and alanine for the second) and makes it understandable why these triplet reiterations do not lie within the coding region of the genes. In the case of new triplet reiterations to be discovered, it should be possible to decide from such considerations whether the triplet is likely to be a codon.

Reiterated hydrophilic amino acids, particularly glutamine, are better tolerated in proteins than hydrophobic or basic amino acids. However, this cannot be the only reason why these amino acids appear to be more commonly reiterated, for the nucleotide sequence of the codons themselves has an important bearing on the frequency of reiterants. The relevant feature is the similarity of the codon to the nucleotide sequence in poly(CAG): for the 3 codons CAG, AGC, and GCA, the frequency of 10 or more reiterations is 30-fold higher per codon than for codons lacking all resemblance to CAG (Table 2). Furthermore, of the 8 codons for glycine and threonine, only the 4 possessing CA, AG, or GC are found in 10-codon reiterants.

Each reiterated unit must consist of a trinucleotide or a multiple of a trinucleotide to generate a homogeneous amino acid reiterant and to prevent the introduction of termination codons by frame-shift. Therefore, where the reiteration mechanism recognizes only one of the dinucleotides of CAG, it must nevertheless act on a trinucleotide or a multiple of it. Even in noncoding regions of the genes for the fragile X mutation and for myotonic dystrophy, the reiterations are of triplets, although simple repeat motifs in noncoding DNA are generally of other than triplets (25).

Evidence for Amino Acid Reiterants Without Function

One may ask what biological function is accomplished by codon reiteration. One possibility is that the cognate amino acid reiterant is important to the function of the protein, as is likely for involucrin (4, 5). Such an interpretation is supported by other examples of homogeneous amino acid reiterants whose encoding nucleotide sequence contains numerous silent substitutions. But codon reiteration is also likely to be a general method of adding new coding region, which is initially without function but which will later be altered by mutation so as to acquire function. If a reiterant, when formed, is not important to the function of the protein, it should at least be compatible with it, and reiterants of hydrophilic residues, whether neutral or acidic, are probably more compatible with different protein structures and functions than are reiterants of hydrophobic or basic residues.

The following four examples support the concept of initially functionless reiteration.

***(i)* Presence or Absence of Polyglutamine in the Corresponding Proteins of Different Species.** One of the earliest demonstrations of variable reiteration is that of interleukin 2. The murine protein contains a sequence of 12 glutamine residues at positions 15–26, a reiterant that is absent from the human protein (26, 27).

***(ii)* Difference in Size of Reiterants at the Same Site in the Corresponding Proteins of Different Species.** Variable reiterants of glutamine are present in TBP of the human (15, 28), the mouse (29), and *Xenopus laevis* (30). Since the amino acid residues flanking the reiterated glutamines are the same in the three widely diverged vertebrate species, the site of reiteration, which is not far from the N terminus, was likely established in a common vertebrate ancestor. Yet the human has 34 glutamine residues at this site, the mouse has 13, and *Xenopus* has 4. In the *Drosophila* protein (31, 32), the N-terminal sequence is not homologous to that of the vertebrates; nevertheless, this region of the protein has two reiterants of 6 and 8 glutamine residues. Therefore, the vertebrates and *Drosophila* have independently introduced glutamine reiterants into their TBP, and within the vertebrates, different lineages have added different numbers of glutamine residues to the same unique site. This requires the repeated operation of a targeted mechanism of genomic change (5, 12, 33). There are no reiterants of glutamine in the TBP gene of *Caenorhabditis elegans*, yeast, *Acanthameba*, *Dictyostelium*, *Plasmodium falciparum*, or various plants (for references, see ref. 13), even though the yeast and acanthameba proteins can function in a mammalian transcription system (34–36). The androgen receptor proteins of the human and the rat (37, 38) possess reiterants of glutamine, arginine, proline, alanine, and glycine at a total of seven sites, but only one of the seven sites has the same number of reiterated residues in the two species (Table 3). For example, the first glutamine reiterant in the human possesses 17 residues, but there is no reiteration at this site in the rat. The glycine reiterant of the human possesses 27 residues, but that

Table 1. Amino acid reiterants and their codons

	CAG-like codons		Non-CAG-like codons	
Amino acid	Codons	No. of reiterants ≥ 10 codons	Codons	No. of reiterants ≥ 10 codons
Q	**CAG·CAG**	15	0	—
	CAA·**CA**A	6		
S	AGC·AGC	5	TCT	0
	TCA·TCA	0	TCC	0
	AGT·AGT	1	TCG	0
A	GCA·GCA	1	0	—
	GCG·GCG	0		
	GCC·GCC	0		
	GCT·GCT	0		
C	TGC·TGC	0	TGT	0
E	GAG·GAG	2	0	—
	GAA·GAA	2		
G	GGA·GGA	1	GGG	0
	GGC·GGC	3	GGT	0
H	CAT·CAT	1	0	—
	CAC·CAC	0		
I	ATC·ATC	0	ATT	0
			ATA	0
K	AAG·AAG	0	AAA	0
L	CTG·CTG	0	TTA	0
			TTG	0
			CTC	0
			CTT	0
			CTA	0
N	AAC·AAC	2	AAT*	6
P	CCA·CCA	0	CCT	0
	CCG·CCG	0	CCC	0
R	AGA·AGA	0	CGA	0
	AGG·AGG	0	CGT	0
	CGC·CGC	0		
	CGG·CGG	0		
T	ACA·ACA	1	ACT	0
	ACC·ACC	1	ACG	0
V	GTA·GTA	0	GTG	0
			GTC	0
			GTT	0
D	0	—	GAT	1
			GAC	0
F	0	—	TTT	0
			TTC	0
M	0	—	ATG	0
W	0	—	TGG	0
Y	0	—	TAT	0
			TAC	0
Total	29	41	32	7

Number of uninterrupted reiterants of 10 or more codons is given for each codon. Generalizations in text on codon reiteration are based on reiterants of this length. Codons are divided into two classes, according to whether or not they resemble the glutamine codon CAG. Nucleotides in boldface are those that occur in reiterated CAG.

*Codon reiterant found only in microorganisms.

Table 2. Nucleotide sequences predisposing to reiteration

Tri- or dinucleotides (in- or out-of-frame)	No. of possible codons	No. of reiterants found	Amino acids reiterated	Reiterants per codon
CAG	3	21	Q, A, S	7.0
CA	9	10	T, H, Q, N	1.1
AG	8	7	S, E, G	0.88
GC	9	3	G	0.33
All remaining codons (no CA, AG, or GC)	32	7	N, D	0.22

The 48 reiterants (≥10 codons) listed in Table 1 are included.

of the rat has only 5. The situation is reversed for some of the other reiterants. Variability in the number of reiterated glutamine residues at corresponding sites in the glucocorticoid receptors of human, mouse, and rat have also been noted (38). There is marked variability of reiterants of glutamine, glycine, and aspartic acid at the 5′ end of the mastermind genes of *Drosophila virilis* and *melanogaster* (18).

(*iii*) Difference in Identity of Reiterated Amino Acid at the Same Site in Corresponding Proteins of Different Species. The C-terminal part of the male sex-determining protein, Sry, is rapidly evolving (39, 40). In *Mus musculus*, this part contains 13 stretches of polyglutamine, but in other species of old world mice and rats, it contains either no reiterant or a reiterant of another amino acid. In one of these species, *Mastomys hildebrantii*, the reiterant consists of multiple alanines, the result of a single nucleotide insertion changing the reading frame from CAG to GCA codons (40).

(*iv*) Absence of Effect of Amino Acid Reiterant by Study of Deletion and of Hybrid Proteins. Deletion of two of the three polyglutamine tracts of the human androgen receptor did not reduce its ability to activate an androgen-sensitive reporter gene (41, 42). Deletion of the entire N terminus of murine interleukin 2 (residues 1–26), including the polyglutamine segment, resulted in an insignificant reduction in the activity of the protein (43). Most differences *in vivo* between the transcriptional activity of TBPs of different species are due to divergences in the sequence of the core or C-terminal region (44, 45), and no effects on the activity of TBP have been localized precisely to the polyglutamine segments of the N-terminal region in vertebrates or *Drosophila.*

A glutamine-rich region that may have begun as a reiterant but has since been modified by nucleotide substitution may be important to the function of the protein. For example, the glutamine-rich regions of the transcription factor SP1 (46) make important contributions to its activity (47) by interacting with a coactivator TBP-associated factor (TAF) (48). However, the glutamine residues themselves are less important in this interaction than adjacent hydrophobic residues (49).

Mechanism of Reiteration

It has been proposed that the mechanism of formation of simple repetitive DNA is replication slippage or slipped-strand mispairing (25). This process may produce reiterations of a single nucleotide or of groups of two, four, five, or seven nucleotides, such as occur in noncoding DNA (refs. 25 and 50 and references therein), and likely would destroy the function

Table 3. Amino acid reiterants at homologous sites in the androgen receptors of two species

Reiterant location (codon no.)		Reiterated amino acid	Size of reiterant (no. of residues)	
Rat	Human		Rat	Human
58	58	Q	1	17
63	79	R	5	1
68	80	Q	2	6
174	189	Q	22	5
373	368	P	4	8
396	394	A	5	5
446	445	G	5	27

Homologous sites may contain reiterants of >4 residues in one or both species. Amino acid sequences of the human DNA-binding domain (residues 556–623) and the human androgen-binding domain (residues 666–918) contain no reiterant and are identical in the two species. The regions containing the reiterants are less well conserved with respect to amino acid replacements.

of a protein by frameshift and premature termination. In addition, slipped-strand mispairing does not favor the reiteration of CAG-like sequences *in vitro* (51). Therefore, the mechanism of codon reiteration should be more precisely targetable than slipped-strand mispairing. The mechanism should not disturb the reading frame, and it must be capable of acting repeatedly at the same site. A final and most interesting property of the codon-reiteration mechanism is that it seems to be able to distinguish between the sense and the antisense strands of the DNA. This is so because among the reiterants listed in Table 1, there are none containing TG (anti-CA) or CT (anti-AG) in the sense strand. This may be interpreted to mean that, although the reiteration mechanism reads either the sense or the antisense strand, the CAG-related sequences specially subject to reiteration must be in the sense strand. One could object that the dinucleotides TG and CT are often found in reiterants encoding hydrophobic amino acids that would be poorly tolerated in proteins. However, selection at the protein level could not explain the fact that there are reiterants of GGA, AGT, and ACC, but not of their antisense trinucleotides, that encode the same three amino acids (glycine, serine, and threonine) but lack CA and AG. Finally, of the six serine codons, all that are reiterated possess at least AG, but the two codons that would, if reiterated, possess the antisense CT are not found among the reiterants of Table 1.

Potential of Homogeneous Codon Reiteration

Codon reiteration is clearly a mechanism for increasing the size of a protein. Owing to the fact that certain codons are much more likely to be reiterated than others, it should be infrequent that the reiterated amino acid will immediately contribute to the function of the protein. CAG reiterants in the involucrin gene are perhaps exceptional because of the particular role of glutamine residues in the protein (5, 6, 52), a role that could explain why long CAG reiterants in coding regions of other genes have led to human disease (7). Homogeneous codon reiterants provide opportunities for evolution by growth of the coding region and mutational modification of the newly added sequence. But it seems necessary that this modification take place soon after generation of long reiterants to prevent ill effects of even the most innocuous homopolymer (polyglutamine).

This investigation was aided by a grant from the National Cancer Institute.

1. Poole, S. J., Kauvar, L. M., Drees, B. & Kornberg, T. (1985) *Cell* **40**, 37–43.
2. Wharton, K. A., Yedvobnick, B., Finnerty, V. G. & Artavanis-Tsakonas, S. (1985) *Cell* **40**, 55–62.
3. Laughon, A., Carroll, S. B., Storfer, F. A., Riley, P. D. & Scott, M. P. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50**, 253–262.
4. Eckert, R. L. & Green, H. (1986) *Cell* **46**, 583–589.
5. Green, H. & Djian, P. (1992) *Mol. Biol. Evol.* **9**, 977–1017.
6. Djian, P., Phillips, M., Easley, K., Huang, E., Simon, M., Rice, R. & Green, H. (1993) *Mol. Biol. Evol.* **10**, 1136–1149.
7. Green, H. (1993) *Cell* **74**, 955–956.
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
9. White, S. H. (1992) *J. Mol. Biol.* **227**, 991–995.
10. Riggins, G. J., Lokey, L. K., Chastain, J. L., Leiner, H. A., Sherman, S. L., Wilkinson, K. D. & Warren, S. T. (1992) *Nat. Genet.* **2**, 186–191.
11. Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. & Wright, F. (1988) *Nucleic Acids Res.* **16**, 8207–8211.
12. Treier, M., Pfeifle, C. & Tautz, D. (1989) *EMBO J.* **8**, 1517–1525.
13. Hancock, J. M. (1993) *Nucleic Acids Res.* **21**, 2823–2830.
14. Ko, C., Smith, C. K., II, & McDonell, M. (1990) *Mol. Biochem. Parasitol.* **41**, 53–64.
15. Hoffmann, A., Sinn, E., Yamamoto, T., Wang, J., Roy, A., Horikoshi, M. & Roeder, R. G. (1990) *Nature (London)* **346**, 387–390.
16. Laurent, B. C., Treitel, M. A. & Carlson, M. (1990) *Mol. Cell. Biol.* **10**, 5616–5625.
17. Muchardt, C. & Yaniv, M. (1993) *EMBO J.* **12**, 4279–4290.
18. Newfield, S. J., Smoller, D. A. & Yedvobnick, B. (1991) *J. Mol. Evol.* **32**, 415–420.
19. Okita, T. W., Cheesbrough, V. & Reeves, C. D. (1985) *J. Biol. Chem.* **260**, 8203–8213.
20. Trumbly, R. J. (1988) *Gene* **73**, 97–111.
21. Caskey, C. T., Pizzuti, A., Fu, Y.-H., Fenwick, R. G., Jr., & Nelson, D. L. (1992) *Science* **256**, 784–789.
22. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-Macdonald, J., de Jong, P. J., Wieringa, B. & Korneluk, R. G. (1992) *Science* **255**, 1253–1255.
23. Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J. & Housman, D. E. (1992) *Cell* **68**, 799–808.
24. Fu, Y.-H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F. & Caskey, C. T. (1992) *Science* **255**, 1256–1258.
25. Levinson, G. & Gutman, G. A. (1987) *Mol. Biol. Evol.* **4**, 203–221.
26. Yokota, T., Arai, N., Lee, F., Rennick, D., Mosmann, T. & Arai, K.-i. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 68–72.
27. Taniguchi, T., Matsui, H., Fujita, T., Takaoka, C., Kashima, N., Yoshimoto, R. & Hamuro, J. (1983) *Nature (London)* **302**, 305–310.
28. Kao, C. C., Lieberman, P. M., Schmidt, M. C., Zhou, Q., Pei, R. & Berk, A. J. (1990) *Nature (London)* **248**, 1646–1650.
29. Tamura, T.-a., Sumita, K., Fujino, I., Aoyama, A., Horikoshi, M., Hoffman, A., Roeder, R. G., Muramatsu, M. & Mikoshiba, K. (1991) *Nucleic Acids Res.* **19**, 3861–3865.
30. Hashimoto, S., Fujita, H., Hasegawa, S., Roeder, R. G. & Horikoshi, M. (1992) *Nucleic Acids Res.* **20**, 3788.
31. Hoey, T., Dynlacht, B. D., Peterson, M. G., Pugh, B. F. & Tjian, R. (1990) *Cell* **61**, 1179–1186.
32. Muhich, M. L., Iida, C. T., Horikoshi, M., Roeder, R. G. & Parker, C. S. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9148–9152.
33. Dover, G. (1982) *Nature (London)* **299**, 111–117.
34. Cavallini, B., Faus, I., Matthes, H., Chipoulet, J. M., Winsor, B., Egly, J. M. & Chambon, P. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9803–9807.
35. Horikoshi, M., Wang, C. K., Fujii, H., Cromlish, J. A., Weil, P. A. & Roeder, R. G. (1989) *Nature (London)* **341**, 299–303.
36. Wong, J.-M., Liu, F. & Bateman, E. (1992) *Gene* **117**, 91–97.
37. Chang, C., Kokontis, J. & Liao, S. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7211–7215.
38. Lubahn, D. B., Joseph, L. R., Sar, M., Tan, J.-a., Higgs, H. N., Larson, R. E., French, F. S. & Wilson, E. M. (1988) *Mol. Endocrinol.* **2**, 1265–1275.
39. Whitfield, L. S., Lovell-Badge, R. & Goodfellow, P. N. (1993) *Nature (London)* **364**, 713–715.
40. Tucker, P. K. & Lundrigan, B. L. (1993) *Nature (London)* **364**, 715–717.
41. Simental, J. A., Sar, M., Lane, M. V., French, F. S. & Wilson, E. M. (1991) *J. Biol. Chem.* **266**, 510–518.
42. Mhatre, A. N., Trifiro, M. A., Kaufman, M., Kazemi-Esfarjani, P., Figlewicz, D., Rouleau, G. & Pinsky, L. (1993) *Nat. Genet.* **5**, 184–188.
43. Zurawski, S. M. & Zurawski, G. (1988) *EMBO J.* **7**, 1061–1069.
44. Cormack, B. P., Strubin, M., Ponticelli, A. S. & Struhl, K. (1991) *Cell* **65**, 341–348.
45. Poon, D., Schroeder, S., Wang, C. K., Yamamoto, T., Horikoshi, M., Roeder, R. G. & Weil, P. A. (1991) *Mol. Cell. Biol.* **11**, 4809–4821.
46. Kadonaga, J. T., Carner, K. R., Masiarz, F. R. & Tjian, R. (1987) *Cell* **51**, 1079–1090.
47. Courey, A. J. & Tjian, R. (1988) *Cell* **55**, 887–898.
48. Hoey, T., Weinzierl, R. O. J., Gill, G., Chen, J.-L., Dynlacht, B. D. & Tjian, R. (1993) *Cell* **72**, 247–260.
49. Gill, G., Pascal, E., Tseng, Z. H. & Tjian, R. (1993) *Proc. Natl. Acad. Sci. USA* **91**, 192–196.
50. Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. (1992) *Genomics* **12**, 241–253.
51. Schlötterer, C. & Tautz, D. (1992) *Nucleic Acids Res.* **20**, 211–215.
52. Simon, M. & Green, H. (1988) *J. Biol. Chem.* **263**, 18093–18098.